INTERNATIONAL APPLICATION AS FILED

明細書

対象音検出方法、信号入力遅延時間検出方法及び音信号処理装置

技術分野

本発明は、検出対象音を検出する対象音検出方法及びそのプログラム、複数のマイクロホンに入力される音信号間の遅延時間を検出する信号入力遅延時間検出方法及びそのプログラム、入力された音信号を処理する音信号処理装置、並びに発話音を検出し、その発話音について音声認識処理を行う音声認識装置に関する。

背景技術

音声は、人間の用いる種々の通信の形態の中でも最も根源的であると同時に、他のどの情報送出方法よりも高速度に情報を送り出すことのできる優れた通信手段である。このようなことから、音声は、古くから現在に至るまで人間の通信手段の根幹を担ってきた。

また、そのような音声を認識するための音声認識技術がある。音声認識とは、その音声に含まれる情報の中で、最も基本的な意味内容に関する情報、つまり音韻情報をコンピュータなどにより抽出し、その抽出内容を判定することである。近年では、計算機プロセッサ技術の飛躍的な発達と、インターネットに代表される高度な情報ネットワークの構築により、様々な分野においてマン・マシンインタフェースとしての音声認識技術の適用が試みられている。

現在の音声認識システムの認識性能は、確率・統計的手法により格段に向上しており、理想的な環境下での音声や接話マイクロホンで収録された近距離音声などでは、非常に高い認識率が得られるようになっている。

ところで、実環境下の音声認識は、学習データと観測データとの間の環境、発話内容などのミスマッチにより、その認識率が劣化する。また、受音系となる接話マイクヘッドセットの装着によりユーザが受ける負担や不快感は大きく、音声認識システム実用化の大きな障害のひとつになっている。

また、S／N比の低下や背景雑音、室内残響の影響などにより認識が困難な遠

隔音声に関し、複数の遠隔マイクロホンを用いた音声認識手法の研究が多くなされている。その代表的なものとして、マイクロホンアレーを用いる手法が挙げられる。この手法では、音源位置検出処理、目的音強調処理、雑音抑制処理、の3つの空間的な信号処理を行なうことができる。このような手法により遠隔音声の音声認識が盛んに研究されている。

しかし、この手法は、正確な話者方向同定処理のために複数のマイクロホンを一定間隔にて固定配置する必要があり、小型化、携帯化が困難であるため、様々な環境・状況下での音声入力への応用が難しく、用途が限定されるという問題がある。

ここで、いつでもどこでも音声入力を可能にするユビキタスな受音系として、①小型・軽量で脱着が容易、②接話マイクとほぼ同等の近距離音声を確保することができる、③接話マイクヘッドセットに比べ、装着時のユーザの負担や不快感を軽減できる、という点で、衣服や眼鏡などに取り付けることができる装着型マイクロホンが期待されている。

本発明は、前述の問題に鑑みてなされたものであり、複数の装着型マイクロホンを用いた環境変動に対してもロバストな受音系の構築を可能にする対象音検出方法、信号入力遅延時間検出方法、音信号処理装置、音声認識装置及プログラムの提供を目的とする。

発明の開示

本発明に係る対象音検出方法は、検出対象音源から出力された検出対象音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記検出対象音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、当該複数のマイクロホンが受音した前記検出対象音を検出することを特徴とする。

また、前記対象音検出方法において、前記周波数を帯域分割して、その分割した帯域毎の前記傾きに基づいて、前記検出対象音を検出することを特徴とする。

また、前記対象音検出方法において、前記帯域毎のそれぞれの傾きが特定の傾

きに集中する傾向が強くなったときに検出対象音を検出することを特徴とする。

また、前記対象音検出方法において、複数のマイクロホンに入力された音信号を所定時間ごとに区切り、各区間の音信号毎に前記クロススペクトルの位相を検出していることを特徴とする。

また、本発明に係る信号入力遅延時間検出方法は、音源から出力された音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出することを特徴とする。

また、前記信号入力遅延時間検出方法において、前記周波数を帯域分割して、その分割した帯域毎の前記傾きに基づいて、前記受音の遅延時間を検出することを特徴とする。

また、前記信号入力遅延時間検出方法において、前記帯域毎のそれぞれの傾きが特定の傾きに集中する傾向が強くなったときに、前記受音の遅延時間を検出することを特徴とする。

また、前記信号入力遅延時間検出方法において、複数のマイクロホンに入力された音信号を所定時間ごとに区切り、各区間の音信号毎に前記クロススペクトルの位相を検出していることを特徴とする。

また、本発明に係る音信号処理装置は、複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、当該複数のマイクロホンが受音した検出対象音源から出力された検出対象音を検出する対象音検出手段と、を備えたことを特徴とする。

また、前記音信号処理装置において、前記傾き検出手段は、前記クロススペクトルの位相の周波数を帯域分割し、分割した帯域毎に傾きを検出しており、前記対象音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、

前記検出対象音を検出することを特徴とする。

また、本発明に係る音信号処理装置は、音源から出力された音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された音を処理する音信号処理装置において、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出する遅延時間検出手段と、前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段と、を備えたことを特徴とする。

また、前記音信号処理装置において、前記傾き検出手段は、前記クロススペクトルの位相を帯域分割し、分割した帯域毎に傾きを検出しており、

また、前記音信号処理装置において、前記遅延時間検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記受音の遅延時間を検出することを特徴とする。

また、本発明に係る音信号処理装置は、検出対象音源から出力された検出対象音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された検出対象音を処理する音信号処理装置において、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記複数のマイクロホン間での前記検出対象音源からの受音の遅延時間を検出する遅延時間検出手段と、前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段と、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記音信号合成手段が合成した合成音信号中の前記検出対象音を検出する対象音検出手段と、を備えたことを特徴とする。

また、前記音信号処理装置おいて、前記傾き検出手段は、前記クロススペクト

ルの位相を帯域分割し、分割した帯域毎に傾きを検出しており、前記遅延時間検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記受音の遅延時間を検出し、前記対象音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記検出対象音を検出することを特徴とする。

また、本発明に係る音声認識装置は、発話源から出力された発話音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された発話音を処理する音声認識装置において、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、当該複数のマイクロホンが受音した前記発話音を検出する発話音検出手段と、前記発話音検出手段が検出した前記発話音について、音声認識処理を行う音声認識処理手段と、を備えたことを特徴とする。

また、前記音声認識装置において、前記傾き検出手段は、前記クロススペクトルの位相の周波数を帯域分割し、分割した帯域毎に傾きを検出しており、前記発話音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記発話音を検出することを特徴とする。

また、本発明に係る音声認識装置は、発話源から出力された発話音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された発話音を処理する音声認識装置において、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記複数のマイクロホン間での前記発話源からの受音の遅延時間を検出する遅延時間検出手段と、前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段と、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記音信号合成手段が合成した合成音信号中の前記発話音を検出する発話音検出手段と、前記発話音検出手段が検出した前記発話音について、音声認識処理を行う音声

認識処理手段と、を備えたことを特徴とする。

また、前記音声認識装置において、前記傾き検出手段は、前記クロススペクトルの位相を帯域分割し、分割した帯域毎に傾きを検出しており、前記遅延時間検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記受音の遅延時間を検出し、前記発話音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記発話音を検出することを特徴とする。

また、本発明に係るプログラムは、検出対象音源から出力された検出対象音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記検出対象音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、当該複数のマイクロホンが受音した前記検出対象音源から出力された検出対象音を検出する処理をコンピュータに実行させることを特徴とする。

また、本発明に係るプログラムは、音源から出力された音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出する処理をコンピュータに実行させることを特徴とする。

ここで、複数のマイクロホンで受音した複数の音信号のクロススペクトルの位相をみた場合、音源と各マイクロホンとの間のそれぞれの距離の差に対応して、その位相の周波数に対する傾きが一定になる。そして、音源と各マイクロホンとの間のそれぞれの距離の差は、複数のマイクロホン間での受音の遅延時間として現れる。さらに、複数のマイクロホンで受音した音声のＳ／Ｎ比が高ければ、そのように傾きが一定となる傾向が顕著になる。本発明はこのような関係を利用したものである。

すなわち、本発明では、複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出

し、その傾きに基づいて、当該複数のマイクロホンが受音した検出対象音や発話音を検出している。なお、検出対象音には、人間が発する発話音の他、物体が発する音も含まれる。

この発明は、複数のマイクロホンで受音した複数の音信号のクロススペクトルの位相をみた場合、音源から各マイクロホンとの距離の差に対応して、その位相の周波数に対する傾きが一定になり、その一方で、複数のマイクロホンで受音した音のS／Nが高ければ、そのように傾きが一定となる傾向が顕著になること、を原理としたものである。

また、本発明では、複数の複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、音源と前記複数のマイクロホンとの間のそれぞれの距離の差に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、前記複数のマイクロホン間での受音の遅延時間を検出している。

この発明は、複数のマイクロホンで受音した複数の音信号のクロススペクトルの位相をみた場合、音源と各マイクロホンとの間のそれぞれの距離の差に対応して、その位相の周波数に対する傾きが一定になり、その一方で、音源と各マイクロホンとの間のそれぞれの距離の差が、複数のマイクロホン間での受音の遅延時間として現れること、を原理とするものである。

また、本発明では、クロススペクトルの位相の周波数を帯域分割し、分割した帯域毎の前記傾きに基づいて処理を行っている。これにより、精度を上げて前記傾きを検出している。

図面の簡単な説明

図１は、本発明の実施の形態の音声信号処理装置を含むシステム全体の構成を示すブロック図である。図２は、本発明の第１の実施の形態の音声信号処理装置の構成を示すブロック図である。図３は、各環境のクロススペクトルの位相を示す特性図である。図４は、クロススペクトルの位相を示す特性図であり、（A）は、音声区間フレームのクロススペクトルの位相を示す特性図であり、（B）は、非音声区間フレームのクロススペクトルの位相を示す特性図である。図５は、

クロススペクトルの位相に基づいて得たヒストグラムを示す特性図であり、（A）は、音声区間フレームのヒストグラムを示す特性図であり、（B）は、非音声区間フレームのヒストグラムを示す特性図である。図６は、音声信号処理装置のヒストグラム等計算部などの構成を示すブロック図である。図７は、第１の実施の形態の音声信号処理装置の効果の説明に用いた特性図である。図８は、本発明の第２の実施の形態の音声信号処理装置の構成を示すブロック図である。図９は、合成信号を生成するためのオーバーラップアッド法の説明に用いた図である。図１０は、第２の実施の形態の音声信号処理装置の効果の説明に用いた特性図である。図１１は、本発明の第３の実施の形態の音声信号処理装置の構成を示すブロック図である。図１２は、音声信号処理装置の音声／非音声判定部の他の構成を示すブロック図である。

発明を実施するための最良の形態

以下、本発明の実施の形態を図面を参照しながら詳細に説明する。この実施の形態は、図１に示すように、２つのマイク１，２で受音した音声信号を処理する音声信号処理装置１０である。ここで、第１及び第２マイク１，２は音源（ユーザ）自体に比較的自由度を持たせた位置に装着可能な装着型マイクである。

図２は、第１の実施の形態の音声信号処理装置１０の構成を示す。図２に示すように、音声信号処理装置１０は、第１及び第２フレーム化部１１，１２と、第１及び第２周波数分析部１３，１４と、クロススペクトル計算部１５と、位相抽出処理部１６と、位相 unwrap 処理部１７と、主計算部３０と、音入力オン／オフ制御部１８とを備えている。また、主計算部３０については、周波数帯域分割部３１と、第１乃至第N傾き計算部32_1～32_Nと、ヒストグラム等計算部３３と、音声／非音声判定部３４とを備えている。以下、各部の処理内容を説明する。

第１及び第２マイク１，２から入力された２ｃｈの音声信号はそれぞれ、第１及び第２フレーム化部１１，１２に入力される。また、第１マイク１から入力された音声信号は、音入力オン／オフ制御部１８に入力される。

第１及び第２フレーム化部１１，１２、第１及び第２周波数分析部１３，１４及びクロススペクトル計算部１５により、第１及び第２マイク１，２から入力された２ｃｈの音声信号のクロススペクトルを算出する。

例えば、第１マイク１と第２マイク２といった複数のマイクで受音した音声信号を時間軸上でみた場合、受音した音声信号間に位相差が生じる。これは、音源から各マイク１，２までの距離の違いにより、音源から各マイク１，２までの音声信号の到達時間に差が生じた結果である。

ここで、第１マイク１と第２マイク２とにより受音した音声信号間の遅延時間を計測し、その計測した遅延時間に基づいて位相を同相化し、その後、第１マイク１と第２マイクとでそれぞれ受音した音声信号を加算して同期加算音声を得る場合を考える。例えば、M.Omologo，P.Svaizerらの文献「"Acoustic event loocalization using a crosspower-spectruum phase based technique", Proc.ICASSP94，pp.274-276，(1994)」に、そのように同期加算音声を得る技術が記載されている。

ここで、２つのマイク１，２で受音した音声信号をそれぞれ$x_1(t)$，$x_2(t)$とし、これら$x_1(t)$，$x_2(t)$をフーリエ変換して得られる周波数関数を$X_1(\omega)$，$X_2(\omega)$とする。ここで、$x_2(t)$は、下記（１）式のように$x_1(t)$の時間移動波形であると仮定する。

$$x_2(t) = x_1(t - t_0) \quad \cdots(1)$$

このように仮定した場合、周波数関数$X_1(\omega)$と周波数関数$X_2(\omega)$との関係は下記（２）式のようになる。

$$X_2(\omega) = e^{j\omega t_0} X_1(\omega) \quad \cdots(2)$$

そして、この周波数関数$X_1(\omega)$と周波数関数$X_2(\omega)$とからクロススペクトル$G_{12}(\omega)$が下記（３）式として得られる。

$$G_{12}(\omega) = X_1(\omega)\, X_2^*(\omega) = X_1(\omega)\, e^{j\omega t_0} X_1^*(\omega) = |X_1|^2 e^{j\omega t_0} \quad \cdots(3)$$

ここで、クロススペクトル$G_{12}(\omega)$の指数項はスペクトル領域のチャネル間の時間遅れに対応する。したがって、周波数関数X_2に遅延項$e^{j\omega t_0}$をかけた$X_2(\omega)\, e^{j\omega t_0}$は、周波数関数$X_1$と同相化され、これにより、$X_1(\omega) + X_2$

$(\omega)\ e^{j\omega\tau_0}$ の逆フーリエ変換をチャネル同期加算音声として扱うことができるようになる。

クロススペクトル計算部１５により、このようなクロススペクトル$G_{12}(\omega)$を得る。

そのため、先ず、第１フレーム化部１１では、後段の第１周波数分析部１３のために、第１マイク１から入力される音声信号をフレーム化（或いはフレーム分割）して、当該第１周波数分析部１３に出力する。また、第２フレーム化部１２では、後段の第２周波数分析部１４のために、第２マイク２から入力される音声信号をフレーム化（或いはフレーム分割）して、当該第２周波数分析部１４に出力する。第１及び第２フレーム化部１１，１２は、所定のサンプル数を１フレームとして、入力されてくる音声信号を次々にフレーム化していく。

例えば、マイク１，２に音声が入力（発話入力）されていない場合には、フレームは、音声の入力されていない非音声区間フレームとなり、マイク１，２に音声が入力されている場合には、フレームは、音声の入力（発話入力）されている音声区間フレームとなる。

第１周波数分析部１３は、第１フレーム化部１１からの音声信号をフーリエ変換して周波数関数$X_1(\omega)$を算出して、後段のクロススペクトル計算部１５に出力する。また、第２周波数分析部１４は、第２フレーム化部１２からの音声信号をフーリエ変換して周波数関数$X_2(\omega)$を算出して、後段のクロススペクトル計算部１５に出力する。ここで、第１及び第２周波数分析部１３，１４は、フレーム毎に音声信号をフーリエ変換する。

クロススペクトル計算部１５は、第１及び第２周波数分析部１３，１４からの周波数関数$X_1(\omega)$、$X_2(\omega)$に基づいて、前記（３）式によりクロススペクトル$G_{12}(\omega)$を算出する。

なお、図３には、１フレームについての音声信号のクロススペクトルの位相を示しており、図３中（A）は自動車内で発した音声について得たクロススペクトルの位相であり、図３中（B）はオフィススペース内で発した音声について得たクロススペクトルの位相であり、図３中（C）は防音室内で発した音声について得たクロススペクトルの位相であり、図３中（D）は歩道（屋外）で発した音声

について得たクロススペクトルの位相である。この図３に示すように、フレーム内で、すなわち局所的に、音源と第１マイク１までの距離と音源と第２マイク２までの距離との差に対応して、クロススペクトルの位相が周波数に対してほぼ一定の傾きを示すことがわかる。すなわち、音源と第１マイク１までの距離と音源と第２マイク２までの距離との差に対応して、クロススペクトルの位相成分が一定の傾きを有している。

また、第１及び第２マイク１，２で受音した音声信号のＳ／Ｎ比が高ければ、そのように傾きが一定となる傾向は顕著になるのである。ここで、第１及び第２マイク１，２が装着型マイクなので、第１及び第２マイク１，２により音声を受音した場合のその音声信号はＳ／Ｎ比が高くなり、このようなことから、明らかに一定の傾きを示すものになっている。

クロススペクトル計算部１５は、このような特性を有するクロススペクトル$G_{12}(\omega)$を位相抽出部１６に出力する。

位相抽出部１６では、クロススペクトル計算部１５からのクロススペクトル$G_{12}(\omega)$から位相を抽出（検出）して、その抽出結果を位相 unwrap 処理部１７に出力する。

位相 unwrap 処理部１７では、位相抽出部１６の位相抽出結果に基づいて、クロススペクトル$G_{12}(\omega)$を unwrap 処理して、主計算部３０の周波数帯域分割部３１に出力する。

周波数帯域分割部３１は、帯域分割（セグメント分割）した位相を第１乃至第Ｎ傾き計算部32_1～32_Nそれぞれに出力する。

ここで、音声の入力されていない非音声区間フレームと音声が入力されている音声区間フレームとで、クロススペクトルの位相成分に大きな違いがある。すなわち、音声区間フレームでは、前述したようにクロススペクトルの位相が周波数に対してほぼ一定の傾きを示すが、非音声区間フレームでは、そのようにはならない。ここで、図４を用いて説明する。

図４はクロススペクトル（CRS）の位相を示しており、図４中（A）は、音声区間フレームのクロススペクトルの位相であり、図４中（B）は、非音声区間フレームのクロススペクトルの位相である。

この図４中（A）と図４中（B）との比較からもかわるように、非音声区間フレームでは、クロススペクトルの位相は、周波数に対して特定のトレンドをもたないのである。すなわち、周波数に対してクロススペクトルの位相が一定の傾きを持つ結果とはならない。これは、ノイズの位相がランダムだからである。

これに対して、音声区間フレームでは、周波数に対してクロススペクトルの位相が一定の傾きをもつようになる。そして、この傾きは、音源から各マイク１，２までの距離の差に対応した大きさになる。

このように、音声の入力されていない非音声区間フレームと音声が入力されている音声区間フレームとでは、クロススペクトルの位相成分に大きな違いがある。

このようなことから、位相の回転が生じた場合にも正確にトレンドを追随するために、周波数帯域分割部３１により、位相成分を小さな周波数セグメントに分割（或いは帯域分割）し、後段の第１乃至第N傾き計算部３２$_1$～３２$_N$で、最小２乗法を適用することでセグメント毎に傾きを計算している。この第１乃至第N傾き計算部３２$_1$～３２$_N$はそれぞれ、算出した傾きをヒストグラム等計算部３３に出力する。

ここで、最小２乗法によりセグメント毎に傾きを求める手法は、公知の技術であり、例えば、『「信号処理」「画像処理」のための入門工学社，高井信勝著，工学社，（２０００）』にその技術が記載されている。

ヒストグラム等計算部３３は、第１乃至第N傾き計算部３２$_1$～３２$_N$が算出した前記傾きについて、ヒストグラムを得る。

図５は、ヒストグラム等計算部３３が得たヒストグラムで、セグメント毎に得た傾きについてのヒストグラムを示している。すなわち、この図５は、位相の傾きの分布を示し、全セグメントに対する、各傾きのセグメント数の割合、すなわち頻度を縦軸にとっている。ここで、図５中（A）は、音声区間フレームについてのヒストグラムを示し、図５中（B）は、非音声区間フレームについてのヒストグラムを示す。

この図５中（A）と図５中（B）との比較からもかわるように、音声区間フレームでは、ヒストグラムに明らかにピーク値があり、すなわち傾きがごく狭い範囲に局在して、これにより、ある傾きについて頻度が高くなっている。すなわち、

帯域毎のそれぞれの傾きが特定の傾きに集中する傾向が強くなっている。一方、非音声区間フレームでは、ヒストグラムが平滑となり、傾きが広い範囲にわたって分布している。

このヒストグラム等計算部３３は、このようなヒストグラム化して得た頻度を音声／非音声判定部３４に出力する。なお、このヒストグラム等計算部３３の処理については後で具体例を説明する。

音声／非音声判定部３４は、ヒストグラム等計算部３３からの前記頻度に基づいて、音声区間と非音声区間とを判定する。例えば、前記頻度の平均値周辺の所定の範囲に含まれる傾きの出現頻度が所定の閾値以上の場合、音声区間と判定し、頻度が所定の閾値未満の場合、非音声区間と判定する。

なお、ここでは、前段の処理がフレーム単位の処理となっているので、当該フレームが、音声区間フレーム又は非音声区間フレームのいずれかであるかを判定する。音声／非音声判定部３４は、その判定結果を音入力オン／オフ制御部１８に出力する。

音入力オン／オフ制御部１８には、第１マイク１からの音声信号が入力されており、音入力オン／オフ制御部１８は、音声／非音声判定部３４の判定結果に基づいて、その第１マイク１からの音声信号の後段への出力をオンとオフとを切り換える。具体的には、音声／非音声判定部３４が音声区間と判定した場合、音入力オン／オフ制御部１８は、オンにして音声信号を後段に出力して、音声／非音声判定部３４が非音声区間と判定した場合、音入力オン／オフ制御部１８は、オフにして音声信号を後段に出力しないようにする。

なお、前段の処理がフレーム単位の処理となっているので、音入力オン／オフ制御部１８は、判定対象のフレームに対応した第１マイク１からの音声信号の部位を単位としてオンとオフとを切り換える。

ヒストグラム等計算部３３の処理の具体例を説明する。図６は、その処理を実現するヒストグラム等計算部３３の構成を示す。

ヒストグラム等計算部３３は、第１乃至第N傾き計算部32_1～32_Nが算出した前記傾きのうちから頻度が高い（最頻度の）傾きを算出する構成として、第１スイッチ３３Ｓ１、第２スイッチ３３Ｓ２及び最頻値計算部３３Ｃを備えている。

これにより、第１スイッチ３３Ｓ１を一定時間オン（閉）にして、第１乃至第Ｎ傾き計算部３２$_1$～３２$_N$が算出した一定時間の前記傾きのデータ（或いはデータベース）３３Ｄ１を作成する。このとき、第２スイッチ３３Ｓ２については、オフ（開）にしておく。そして、データ３３Ｄ１を作成したら、第２スイッチ３３Ｓ２をオン（閉）にして、そのデータ３３Ｄ１を最頻値計算部３３Ｃに出力する。

最頻値計算部３３Ｃでは、データ３３Ｄ１から前記図５に示すような前記傾きについてのヒストグラムを作成して、そのヒストグラム中の最頻度の傾き（以下、最頻傾きという。）τ０を算出する。なお、最頻度の傾きを算出するようにしてもよいが、平均値の傾きτ０を算出したり、或いは最頻度の傾きと傾きの平均値とを組み合わせた傾きτ０を算出するようにしてもよい。これにより、各帯域の傾きが特定の傾きに集中する傾向が強くなったとき、当該特定の傾きの値そのもの或いはそれに近い傾きの値を得ることができる。なお、本実施の形態では、最頻値計算部３３Ｃが最頻傾きτ０を算出しているものとする。

そして、最頻値計算部３３Ｃは、算出した最頻傾きτ０を前記音声／非音声判定部３４に出力する。ここで、最頻傾きτ０をデータ３３Ｄ２として前記音声／非音声判定部３４に出力する。

以上がヒストグラム等計算部３３の処理の具体例である。

前記音声／非音声判定部３４では、ヒストグラム等計算部３３からの最頻傾きτ０に基づいて、音声区間と非音声区間とを判定する。

なお、先の説明では、音声／非音声判定部３４がヒストグラム等計算部３３からの前記頻度に基づいて音声区間と非音声区間とを判定する場合について説明した。ここでは、音声／非音声判定部３４は、ヒストグラム等計算部３３からの最頻傾きτ０と第１乃至第Ｎ傾き計算部３２$_1$～３２$_N$が算出した前記傾き（各帯域の傾き）τｉに基づいて、音声区間と非音声区間とを判定しており、これに対応して、音声／非音声判定部３４に、第１乃至第Ｎ傾き計算部３２$_1$～３２$_N$が算出した前記傾きが入力されるようになっている。

すなわち、音声／非音声判定部３４は、第１乃至第Ｎ傾き計算部３２$_1$～３２$_N$が算出した前記傾きτｉと最頻傾きτ０とを下記（４）式により比較する。

$$|\tau i - \tau 0| < \delta \quad \cdots (4)$$

ここで、δは判定用の閾値（傾き閾値）である。

音声／非音声判定部３４は、この（４）式の条件が満たされていることが所定の割合を超えた場合（ＹＥＳ）、音声区間と判定し、そうでない場合（ＮＯ）、非音声区間と判定する。そして、音声／非音声判定部３４は、その判定結果を音入力オン／オフ制御部１８に出力する。

以上のように構成した音声信号処理装置１０の一連の動作は次のようになる。

先ず、第１及び第２フレーム化部１１，１２、第１及び第２周波数分析部１３，１４及びクロススペクトル計算部１５が、第１及び第２マイク１，２から入力された２ｃｈの音声信号のクロススペクトル$G_{12}(\omega)$を算出する。

そして、位相抽出部１６、位相 unwrap 処理部１７及び周波数帯域分割部３１が、そのように算出したクロススペクトル$G_{12}(\omega)$の位相を帯域分割（セグメント分割）して、第１乃至第Ｎ傾き計算部32_1～32_Nが、帯域毎（セグメント毎）の位相の傾きを算出する。

そして、ヒストグラム等計算部３３が、第１乃至第Ｎ傾き計算部32_1～32_Nそれぞれが算出した前記帯域毎（セグメント毎）の傾きからヒストグラムを生成して、音声／非音声判定部３４が、そのヒストグラムから得られる頻度と最頻傾きτ_0に基づいて、音声区間と非音声区間とを判定する。この判定結果に基づいて、音入力オン／オフ制御部１８では、第１マイク１からの音声信号の後段への出力をオンとオフとを切り換える。具体的には、音声／非音声判定部３４が音声区間と判定した場合、音入力オン／オフ制御部１８は、オンにして音声信号を後段に出力して、音声／非音声判定部３４が非音声区間と判定した場合、音入力オン／オフ制御部１８は、オフにして音声信号を後段に出力しないようにする。

このように、音声信号処理装置１０は、第１マイク１，２が受音した音声中の発話区間（音声区間）を検出することができる。

例えば、第１マイク１，２と音声アプリケーションとの間にこのような音声信号処理装置１０を備えることで、音声アプリケーションは、確実に発話区間についての処理を行うことができる。ここで、音声アプリケーションとしては、音声認識システム、放送システム、携帯電話及びトランシーバが挙げられる。例えば、音声アプリケーションが音声認識システムであるとすれば、音声認識システムは、

音声信号処理装置１０が出力する発話区間の音声信号に基づいて音声認識することができるようになる。

次に効果を説明する。

前述したように、第１及び第２マイク１，２に入力された音信号間のクロススペクトルの位相を検出し、その検出したクロススペクトルの位相の周波数に対する傾きに基づいて、当該複数のマイクロホンが受音した音声信号中の発話区間を検出している。すなわち、音声が入力（発話入力）されていない音声信号と音声が入力（発話入力）されている音声信号とをクロススペクトルでみた場合に、そのクロススペクトルの位相成分に大きな違いがあることを利用して、当該複数のマイクロホンが受音した音声信号中の発話区間を検出している。

具体的には、クロススペクトルの位相を帯域分割（セグメント分割）し、帯域毎（セグメント毎）の位相の傾きからヒストグラムを生成し、そのヒストグラムから頻度（具体的には最頻値）を得て、その頻度に基づいて、発話区間を検出している。

これにより、精度よく発話区間を検出することができる。そして、このように音声信号処理装置１０が検出した発話区間の音声信号を利用することにより、音声認識システムでは、高認識率、低誤認識率の音声認識が可能になり、また、携帯電話やトランシーバでは、信頼性の高いハンズフリー半二重通信が可能になり、放送システムでは、通信システムの送信電力低減が可能になる。

また、マイクの取り付け位置等の環境の変化や、話者の移動や姿勢の変化等の音源の移動に対しても、ロバストな音声入力を実現することができる。

前述したように、クロススペクトルの位相の周波数に対する傾きは、音源と第１マイク１までの距離と音源と第２マイク２までの距離との差に対応して変化する値になっている。これにより、例えば、音源に対する第１及び第２マイク１，２の取り付け位置を変更した場合、クロススペクトルの位相の周波数に対する傾きはその位置の変更に対応して変化するようになる。その一方で、前述したように、クロススペクトルの位相を帯域分割（セグメント分割）し、帯域毎（セグメント毎）の位相の傾きからヒストグラムを生成し、そのヒストグラムから頻度（具体的には最頻度）を得て、その頻度に基づいて、発話区間を検出している。

すなわち、クロススペクトルの位相の傾きの大きさ自体に拠ることなく、つまり、音源とマイク１，２との間の距離に左右されることなく、最終的に、発話区間の検出を行っている。よって、音源に対する第１及び第２マイク１，２の取り付け位置を変更した場合でも、発話区間の検出結果への影響はない。

このようなことから、マイクの取り付け位置等の環境の変化や、話者の移動や姿勢の変化等の音源の移動に対しても、ロバストな音声入力を実現することができる。すなわち、マイクの位置の自由度を高くしつつ、ロバストな音声入力を実現することができる。

以上のように、小型・軽量で脱着が容易であり、接話マイクとほぼ同等の近距離音声を確保することができ、接話マイクヘッドセットに比べ、装着時のユーザの負担や不快感を軽減できる装着型マイクを用いることを前提としつつも、前述した種々の効果を得ることができる。

（実施例（第１の実施の形態））

本発明を適用したシステムにより音声の発話区間の検出を行った。各文章間に１秒程度の無発話区間を含む合計４０文をサンプルの使用音声とした。実験環境は、防音室内、自動車内、オフィススペース内及び歩道上といった環境とした。評価方法は、①無音声区間フレームを音声区間フレームであると誤判別した場合、②発話区間の始端・終端において、発話区間を無発話区間であると誤判別した場合、このような①や②に該当する場合のフレームをエラーフレームとした。また、比較対象（従来例）として、平均ゼロ交差回数と対数パワーとを変数としたフィッシャーの線形判別関数による手法を用いた。

図７は、その結果を示す。この図７は、総フレームに対するエラーフレームの割合の百分率（発話区間誤検出率）を示す。図７中、ＬＤＦの値は、前記線形判別関数による手法の値であり、ＣＲＳの値はクロススペクトルを用いた手法（本発明）の値である。

この図７に示すように、防音室内やオフィススペース内においては、発話区間誤検出率の結果に、平均ゼロ交差回数と対数パワーによる方法と本発明による手法とで大きな差はみられない。しかし、自動車内や歩道では、発話区間誤検出率

の結果が本発明による手法により改善される結果を示すようになった。このように、本発明は、特に雑音環境下において有効に作用する。

次に第２の実施の形態を説明する。

図８は、この第２の実施の形態の音声信号処理装置１０の構成を示す。この第２の実施の形態では、第１マイク１と第２マイク２とで受音した音声信号を合成して後段の音声アプリケーションに出力する構成になっている。このため、この第２の実施の形態では、遅延処理部５１と波形合成部５２とを備え、遅延処理部５１で第２マイク２からの音声信号を遅延させて波形合成部５２に出力して、波形合成部５２で、遅延処理部５１で遅延されて入力された第２マイク２の音声信号と第１マイク１からの音声信号とを合成して出力している。

第１マイク１と第２マイク２といった複数のマイクで受音した音声信号間には、音源から各マイク１，２までの距離の違いに起因する位相差が生じる。このようなことから、第１マイク１と第２マイク２といった複数のマイクで受音した音声信号を合成しようとする場合には、音源から各マイク１，２までの音声信号の到達時間差を補正し、位相を同相化したのちに音声信号を加算する、という遅延和処理が必要になる。このようなことから、前述したように、第２の実施の形態では、遅延処理部５１と波形合成部５２とを備えている。

そして、前述の第１の実施の形態では（図６参照）では、最頻値計算部３３Ｃがヒストグラムから最頻傾き$\tau 0$を算出しているが、第２の実施の形態では、そのような最頻傾き$\tau 0$に基づいて、遅延処理部５１で遅延処理しているのである。以下に具体的に説明する。

前記図３や図４中（Ａ）に示すように音声区間ではクロススペクトルの位相成分が一定の傾きを有するが、この傾きは、第１マイク１と第２マイク２とのチャネル間の遅延時間を示すものとなる。

このような関係を利用して、遅延処理部５１では、ヒストグラム等計算部３３が算出した前記最頻傾き$\tau 0$に基づいて、遅延処理している。具体的には、図６に示すように、最頻値計算部３３Ｃから遅延処理部５１に最頻傾き$\tau 0$が出力されており、遅延処理部５１は、入力されたこの最頻傾き$\tau 0$に基づいて遅延処理している。

$$\tau 0 = x / n = 2\pi \cdot n_0 / N \quad [\text{rad/point}] \quad \cdots (5)$$

ここで、x，nの単位はそれぞれラジアン、周波数ポイント（point）であり、Nは、ＦＦＴポイント数であり、n_0は遅延サンプリングポイント数である。

この関係から、下記（6）式として、最頻傾き$\tau 0$を変数とした遅延サンプリングポイント数n_0を得ることができる。

$$n_0 = \tau 0 / (2\pi / N) \quad [\text{point}] \quad \cdots (6)$$

そして、この遅延サンプリングポイント数n_0を用いて、下記（7）式により、遅延時間t_0を得ることができる。

$$t_0 = n_0 / F_s \quad \cdots (7)$$

ここで、F_sは、サンプリング周波数であり、例えば１６ｋＨｚである。

遅延処理部５１では、このようにして得た遅延時間t_0に基づいて、入力される第２マイク２の音声信号を遅延して、波形合成部５２に出力する。

波形合成部５２は、遅延処理部５１で遅延されて入力された第２マイク２の音声信号と第１マイク１からの音声信号とを合成して出力する。

なお、音声信号の合成信号を次のようにして得ることもできる。

前述したように、周波数関数X_2に遅延項$e^{j\omega t_0}$をかけた$X_2(\omega)\ e^{j\omega t_0}$は、周波数関数$X_1$と同相化され、これにより、$X_1(\omega) + X_2(\omega)\ e^{j\omega t_0}$の逆フーリエ変換をチャネル同期加算音声として扱うことができる。この関係を利用して、音声信号の合成信号を得る。

すなわち、先ず遅延時間t_0を用いることで、下記（8）式により、周波数軸上でチャネル同期加算音声$X_1(\omega) + X_2(\omega)\ e^{j\omega t_0}$を得る。ここで、遅延時間$t_0$は、前記（6）式及び（7）式に示すように最頻傾き$\tau 0$を変数とする値である。

$$X_1(\omega) + X_2(\omega)\ e^{j\omega t_0} = \{Re[X_1(\omega)] + j\ Im[X_1(\omega)]\} + \{Re[X_2(\omega)](\cos\omega t_0 + j\sin\omega t_0) + j\ Im[X_2(\omega)](\cos\omega t_0 + j\sin\omega t_0)\} \quad \cdots (8)$$

ここで、チャンネル同期音声スペクトルは、実部、虚部にそれぞれ

$$Re : Re[X_2(\omega)]\cos\omega t_0 - Im[X_2(\omega)]\sin\omega t_0 + Re[X_1(\omega)]$$

$$Im : Re[X_2(\omega)]\sin\omega t_0 + Im[X_2(\omega)]\cos\omega t_0 + Re[X_1(\omega)]$$

を持つ複素スペクトルになる。この処理をフレーム毎に施し、それぞれのフレーム毎にＩＦＦＴ（インバースＦＦＴ）をし、同期加算音声のフレーム列を得る。

そして、そのようにして得たフレーム列にオーバーラップアッド法（Overlap-add method）を適用して同期加算音声、すなわち第１マイク１の音声信号と第２マイク２の音声信号との合成信号を得る。

ここで、オーバーラップアッド法とは、図９に示すように、入力データ列$s_n(t)$を重ね合わせながら加算する方法である。ここで、$s_n(t)$はｎ番目の合成音声波形フレームを示す。また、図中Ｌは定数である。

以上のように音声信号処理装置１０を構成することで、遅延処理部５１が第２マイク２からの音声信号を遅延させて波形合成部５２に出力して、波形合成部５２が、遅延処理部５１により遅延されて入力された第２マイク２からの音声信号と第１マイク１からの音声信号とを合成して出力する。

これによる効果は次のようになる。

前述の第１の実施の形態で説明したように、クロススペクトルの位相の周波数に対する傾きは、音源と第１マイク１までの距離と音源と第２マイク２までの距離との差に対応して変化する値である。このようなクロススペクトルの位相の周波数に対する傾きから前記遅延時間を推定している。そして、実際に推定の際に用いる値を、最頻傾き$\tau 0$としている。このように最頻傾き$\tau 0$を用いて、遅延時間を推定しているので、精度を高くして遅延時間の推定を行うことができる。

そして、このような遅延時間に基づいて、第１マイクと第２マイクとの音声信号を合成することで、高品質の合成音声信号を提供することができる。例えば、このような合成音声信号を利用した場合、音声認識システムでは、高認識率、低誤認識率の音声認識が可能になり、また、携帯電話やトランシーバでは、高品質の音声による通話が可能になり、放送システムでは、高品質の放送や録音が可能になる。

また、遅延時間を推定に用いる前記傾きを、最頻傾き$\tau 0$とした結果、前述の第１の実施の形態と同様に、マイクの取り付け位置等の環境の変化や、話者の移動や姿勢の変化等の音源の移動に対しても、ロバストな音声入力を実現すること

ができる。すなわち、マイクの位置の自由度を高くしつつ、ロバストな音声入力を実現することができる。

以上のように、小型・軽量で脱着が容易であり、接話マイクとほぼ同等の近距離音声を確保することができ、接話マイクヘッドセットに比べ、装着時のユーザの負担や不快感を軽減できる装着型マイクを用いることを前提としつつも、前述した種々の効果を得ることができる。

（実施例（第２の実施の形態））

本発明を適用したシステムにより生成した同期加算音声（合成音声信号）を用いて、音響モデルによる音声認識の実験をした。

音響モデルによる音声認識実験では、先ず、同期加算音声による学習データにより、音響モデルを作成した。作成した音響モデルは次のようになる。

①収録環境毎に作成した４種類の収録環境依存型HMM（hidden Markov model）

②すべて環境の収録音声により学習した収録環境非依存型HMM

ここで、前記収録環境とは、前記防音室内、自動車内、オフィススペース内及び歩道上である。

そして、作成した音響モデルを用いて、音声認識実験を行った。

認識タスクは連続音声認識であり、評価用データ（評価用音声）は学習時と異なる音声としている。図１０は、その音声認識実験で得た認識率の結果を示す。ここで、比較対象（従来例）として、第１マイク１と第２マイク２とからの単チャンネル音声による認識率の結果も示す。例えば、第１マイク１は眼鏡マイクであり、第２マイク２は胸元マイクである。ここで、眼鏡マイクとは、眼鏡のフレームに装着したマイクである。

この図１０に示すように、車内以外の、 防音室内、歩道上及びすべての環境で、本発明により得た同期加算音声による認識率が、単チャネル音声の認識率を上回る結果となっている。これにより、実環境においても、本発明を適用したシステムが生成した同期加算音声が高品質であることがわかる。

次に第３の実施の形態を説明する。

図１１は、この第３の実施の形態の音声信号処理装置１０の構成を示す。この第２の実施の形態の音声信号処理装置１０は、前述の第１の実施の形態の音声信号処理装置１０の構成と、第２の実施の形態の音声信号処理装置１０の構成とを組み合わせた構成になっている。すなわち、第３の実施の形態の音声信号処理装置１０は、音声／非音声判定部３４、遅延処理部５１、波形合成部５２及び音声入力オン／オフ制御部１８を同時に備えている。

このように構成することで、第３の実施の形態の音声信号処理装置１０は以下のように動作する。なお、特に言及しない部分については、前述の第１の実施の形態の音声信号処理装置１０や第２の実施の形態の音声信号処理装置１０と同様に動作するものとする。

遅延処理部５１が、ヒストグラム等計算部３３（最頻値計算部３３Ｃ）が算出した最頻傾きτ０に基づいて、第２マイク２の音声信号を遅延し、波形合成部５２が、遅延処理部５１で遅延されて入力された第２マイク２からの音声信号と第１マイク１からの音声信号とを合成して、合成音声信号を音入力オン／オフ制御部１８に出力する。

一方、音声／非音声判定部３４が、ヒストグラム等計算部３３が得た頻度に基づいて、音声区間と非音声区間とを判定し、音入力オン／オフ制御部１８では、その判定結果に基づいて、波形合成部５２から出力される音声信号（同期加算音声信号）の出力をオン又はオフする。

このように構成することで、第３の実施の形態の音声信号処理装置１０は、前述の第１の実施の形態の音声信号処理装置１０が有する効果と、第２の実施の形態の音声信号処理装置１０が有する効果とを発揮することができる。

すなわち、高品質の合成音声信号を生成するとともに、その合成音声信号中の発話区間を精度よく検出することができる。さらに、マイクの取り付け位置等の環境の変化や、話者の移動や姿勢の変化等の音源の移動に対しても、ロバストな音声入力を実現することができる。すなわち、マイクの位置の自由度を高くしつつ、ロバストな音声入力を実現することができる。

以上、本発明の実施の形態について説明した。しかし、本発明は、前述の実施の形態として実現されることに限定されるものではない。

例えば、図１２に示すように、前記音声／非音声判定部３４が、第１乃至第N傾き計算部32_1～32_Nが算出した前記傾きτiと最頻傾き$\tau 0$とを下記（９）式により比較する。

$$|\tau i - \tau 0| < \alpha\sigma \quad \cdots(9)$$

ここで、αは係数であり、σは前記判定用の閾値（傾き閾値）δに物理的に内包される値である。例えば、δと$\alpha\sigma$とを用意した意味は、δを固定値とし、$\alpha\sigma$をリアルタイム学習により随時更新する変数とし、これにより、各値による音声区間の検出の効果の違いを区別するためである。

$\alpha\sigma$のσを更新することで、静粛な環境では、音声区間判定条件を厳しくし、より非音声区間の誤判定を防止することができる。すなわち、バックグランドノイズのある環境では判定条件を甘くすることで、音声区間を安定して検出することが可能になる。仮に、バックグランドノイズのある環境にもかかわらず静粛環境のσを用いてしまうと、この場合固定値のδを用いることと等価となるが、この場合には、ノイズと音声とが重なっているようなとき、音声区間が棄却されてしまうおそれがある。

すなわち、固定値としてのδは、その値を設定した条件に近い環境での音声区間を検出に用いるときに当該音声区間の検出に有効に作用し、変数である$\alpha\sigma$は、環境の変化に対し動的に対応するシステムに用いるときに音声区間の検出に有効に作用する。

また、係数αを変更することでも、判定を厳しくしたり、甘くしたりすることもできる。

また、前述の実施の形態では、前記帯域毎の傾きをヒストグラム化することで、帯域毎のそれぞれの傾きが特定の傾きに集中する傾向をみている。しかし、他の手法により、帯域毎のそれぞれの傾きが特定の傾きに集中する傾向をみるようにしてもよい。

また、前述の実施形態では、検出対象音が人間が発する発話音である場合を説明したが、検出対象音は、人間以外の物体が発する音でもよい。

また、前述の実施の形態の説明において、第１及び第２フレーム化部１１，１２、第１及び第２周波数分析部１３，１４及びクロススペクトル計算部１５が、

複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段を実現しており、位相抽出部１６、位相unwrap処理部１７、周波数帯域分割部３１及び第１乃至第N傾き計算部32_1～32_Nが、前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段を実現しており、ヒストグラム等計算部３３及び音声／非音声判定部３４が、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、当該複数のマイクロホンが受音した発話音の発話区間を検出する発話音検出手段を実現している。

また、ヒストグラム等計算部３３及び遅延処理部５１が、前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記複数のマイクロホン間での受音の遅延時間を検出する遅延時間検出手段を実現しており、波形合成部５２が、前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段を実現している。

また、前述の実施形態の音声信号処理装置１０を音声認識装置に適用することができる。この場合、音声認識装置は、前述したような音声信号処理装置１０の構成に加えて、音声信号処理装置１０が検出した発話区間の音声信号（発話音）について音声認識処理をする音声認識処理手段を備える。

ここで、音声認識技術としては、例えば、旭化成株式会社が提供する音声認識技術「ＶＯＲＥＲＯ」（商標）（http://www.asahi-kasei.co.jp/vorero/jp/vorero/feature.html参照）等があり、このような音声認識技術の用いた音声認識装置に適用することもできる。

また、前述の実施形態の音声信号処理装置１０をコンピュータで実現することができる。そして、前述したような音声信号処理装置１０の処理内容をコンピュータが所定のプログラムにより実現する。この場合、プログラムは、検出対象音源から出力された検出対象音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記検出対象音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、当該複数のマイクロホンが受音した前記検出対象音源から出力された検

出対象音を検出する処理をコンピュータに実行させるプログラムになる。又は、プログラムは、音源から出力された音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出する処理をコンピュータに実行させるプログラムになる。

産業上の利用の可能性

本発明によれば、装着型マイクロホンを用いた環境変動に対してもロバストな受音系の構築を可能にすることができる。

請求の範囲

1．　検出対象音源から出力された検出対象音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記検出対象音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、当該複数のマイクロホンが受音した前記検出対象音を検出することを特徴とする対象音検出方法。

2．　前記周波数を帯域分割して、その分割した帯域毎の前記傾きに基づいて、前記検出対象音を検出することを特徴とする請求の範囲第１項記載の対象音検出方法。

3．　前記帯域毎のそれぞれの傾きが特定の傾きに集中する傾向が強くなったときに検出対象音を検出することを特徴とする請求の範囲第２項記載の対象音検出方法。

4．　複数のマイクロホンに入力された音信号を所定時間ごとに区切り、各区間の音信号毎に前記クロススペクトルの位相を検出していることを特徴とする請求の範囲第１乃至３項のいずれかに記載の対象音検出方法。

5．　音源から出力された音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出することを特徴とする信号入力遅延時間検出方法。

6．　前記周波数を帯域分割して、その分割した帯域毎の前記傾きに基づいて、前記受音の遅延時間を検出することを特徴とする請求の範囲第５項記載の信号入力遅延時間検出方法。

7．　前記帯域毎のそれぞれの傾きが特定の傾きに集中する傾向が強くなったときに、前記受音の遅延時間を検出することを特徴とする請求の範囲第６項記載の信号入力遅延時間検出方法。

8．　複数のマイクロホンに入力された音信号を所定時間ごとに区切り、各区間の音信号毎に前記クロススペクトルの位相を検出していることを特徴とする請求の範囲第５乃至７項のいずれかに記載の信号入力遅延時間検出方法。

9．　複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、

前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、

前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、当該複数のマイクロホンが受音した検出対象音源から出力された検出対象音を検出する対象音検出手段と、

を備えたことを特徴とする音信号処理装置。

１０．　前記傾き検出手段は、前記クロススペクトルの位相の周波数を帯域分割し、分割した帯域毎に傾きを検出しており、

前記対象音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記検出対象音を検出することを特徴とする請求の範囲第９項記載の音信号処理装置。

１１．　音源から出力された音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された音を処理する音信号処理装置において、

前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、

前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、

前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出する遅延時間検出手段と、

前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段と、

を備えたことを特徴とする音信号処理装置。

１２．　前記傾き検出手段は、前記クロススペクトルの位相を帯域分割し、分割

した帯域毎に傾きを検出しており、

前記遅延時間検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記受音の遅延時間を検出することを特徴とする請求の範囲第１１項記載の音信号処理装置。

１３．　検出対象音源から出力された検出対象音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された検出対象音を処理する音信号処理装置において、

前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、

前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、

前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記複数のマイクロホン間での前記検出対象音源からの受音の遅延時間を検出する遅延時間検出手段と、

前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段と、

前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記音信号合成手段が合成した合成音信号中の前記検出対象音を検出する対象音検出手段と、

を備えたことを特徴とする音信号処理装置。

１４．　前記傾き検出手段は、前記クロススペクトルの位相を帯域分割し、分割した帯域毎に傾きを検出しており、前記遅延時間検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記受音の遅延時間を検出し、前記対象音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記検出対象音を検出することを特徴とする請求の範囲第１３項記載の音声信号処理装置。

１５．　発話源から出力された発話音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された発話音を処理する音声認識装置において、

前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検

出するクロススペクトル位相検出手段と、

前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、

前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、当該複数のマイクロホンが受音した前記発話音を検出する発話音検出手段と、

前記発話音検出手段が検出した前記発話音について、音声認識処理を行う音声認識処理手段と、

を備えたことを特徴とする音声認識装置。

１６．　前記傾き検出手段は、前記クロススペクトルの位相の周波数を帯域分割し、分割した帯域毎に傾きを検出しており、

前記発話音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記発話音を検出することを特徴とする請求の範囲第１５項記載の音声認識装置。

１７．　発話源から出力された発話音が複数のマイクロホンに入力され、前記複数のマイクロホンに入力された発話音を処理する音声認識装置において、

前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出するクロススペクトル位相検出手段と、

前記クロススペクトル位相検出手段が検出したクロススペクトルの位相の周波数に対する傾きを検出する傾き検出手段と、

前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記複数のマイクロホン間での前記発話源からの受音の遅延時間を検出する遅延時間検出手段と、

前記遅延時間検出手段が検出した遅延時間に基づいて、前記複数のマイクロホンに入力された音信号同士を合成する音信号合成手段と、

前記傾き検出手段が検出した前記周波数に対する傾きに基づいて、前記音信号合成手段が合成した合成音信号中の前記発話音を検出する発話音検出手段と、

前記発話音検出手段が検出した前記発話音について、音声認識処理を行う音声認識処理手段と、

を備えたことを特徴とする音声認識装置。

１８．　前記傾き検出手段は、前記クロススペクトルの位相を帯域分割し、分割した帯域毎に傾きを検出しており、前記遅延時間検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記受音の遅延時間を検出し、前記発話音検出手段は、前記傾き検出手段が検出した前記帯域毎の傾きに基づいて、前記発話音を検出することを特徴とする請求の範囲第１７項記載の音声認識装置。

１９．　検出対象音源から出力された検出対象音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記検出対象音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、当該複数のマイクロホンが受音した前記検出対象音源から出力された検出対象音を検出する処理をコンピュータに実行させることを特徴とするプログラム。

２０．　音源から出力された音が複数のマイクロホンに入力されており、前記複数のマイクロホンに入力された音信号間のクロススペクトルの位相を検出し、前記音源と前記複数のマイクロホンとの間のそれぞれの距離に起因して発生する前記クロススペクトルの位相の周波数に対する傾きを検出し、その傾きに基づいて、前記複数のマイクロホン間での前記音源からの受音の遅延時間を検出する処理をコンピュータに実行させることを特徴とするプログラム。

要約書

装着型マイクロホンを用いた環境変動に対してもロバストな受音系の構築を可能にする。

音声信号処理装置１０は、マイク１，２に入力された音信号間のクロススペクトルの位相を検出する第１及び第２フレーム化部１１，１２、第１及び第２周波数分析部１３，１４及びクロススペクトル計算部１５と、クロススペクトル計算部１５が検出したクロススペクトルの位相の周波数に対する傾きを検出する位相抽出部１６、位相 unwrap 処理部１７、周波数大域分割部３１及び第１乃至第N傾き計算部32_1～32_Nと、第１乃至第N傾き計算部32_1～32_Nが検出した前記周波数に対する傾きに基づいて、マイク１，２が受音した発話の発話区間を検出するヒストグラム等計算部３３及び音声／非音声判定部３４とを備える。

図１

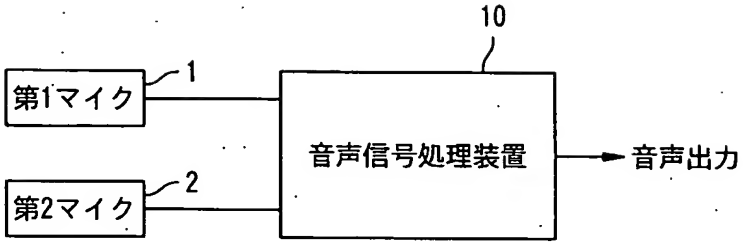


図2

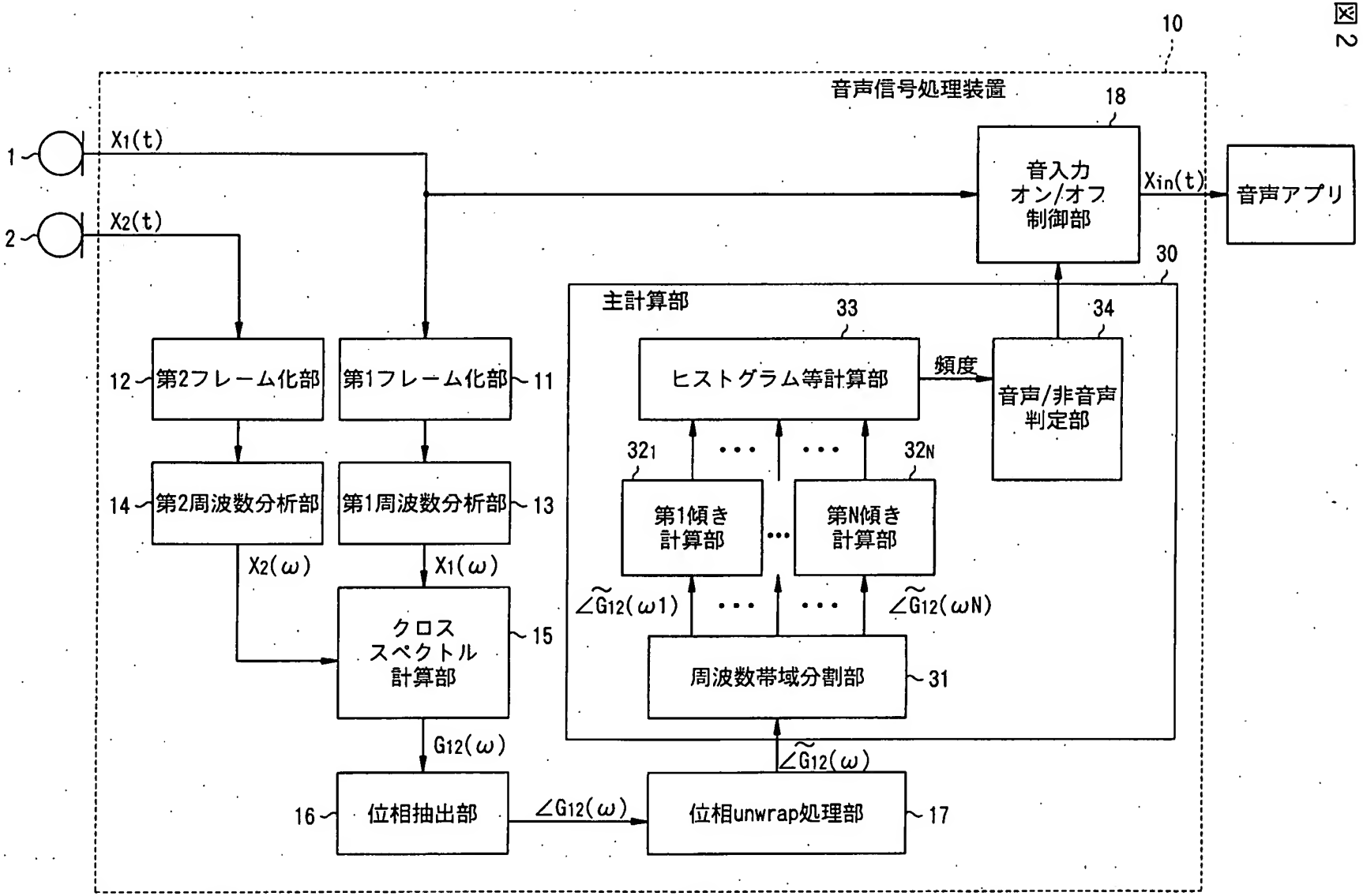

10
音声信号処理装置
18
音入力
オン/オフ
制御部
Xin(t)
音声アプリ
1
X1(t)
2
X2(t)
30
主計算部
33
34
ヒストグラム等計算部
頻度
音声/非音声
判定部
12
第2フレーム化部
第1フレーム化部
11
321
32N
第1傾き
計算部
第N傾き
計算部
14
第2周波数分析部
第1周波数分析部
13
X2(ω)
X1(ω)
∠G̃12(ω1)
∠G̃12(ωN)
クロス
スペクトル
計算部
15
周波数帯域分割部
31
G12(ω)
∠G̃12(ω)
16
位相抽出部
∠G12(ω)
位相unwrap処理部
17

図3

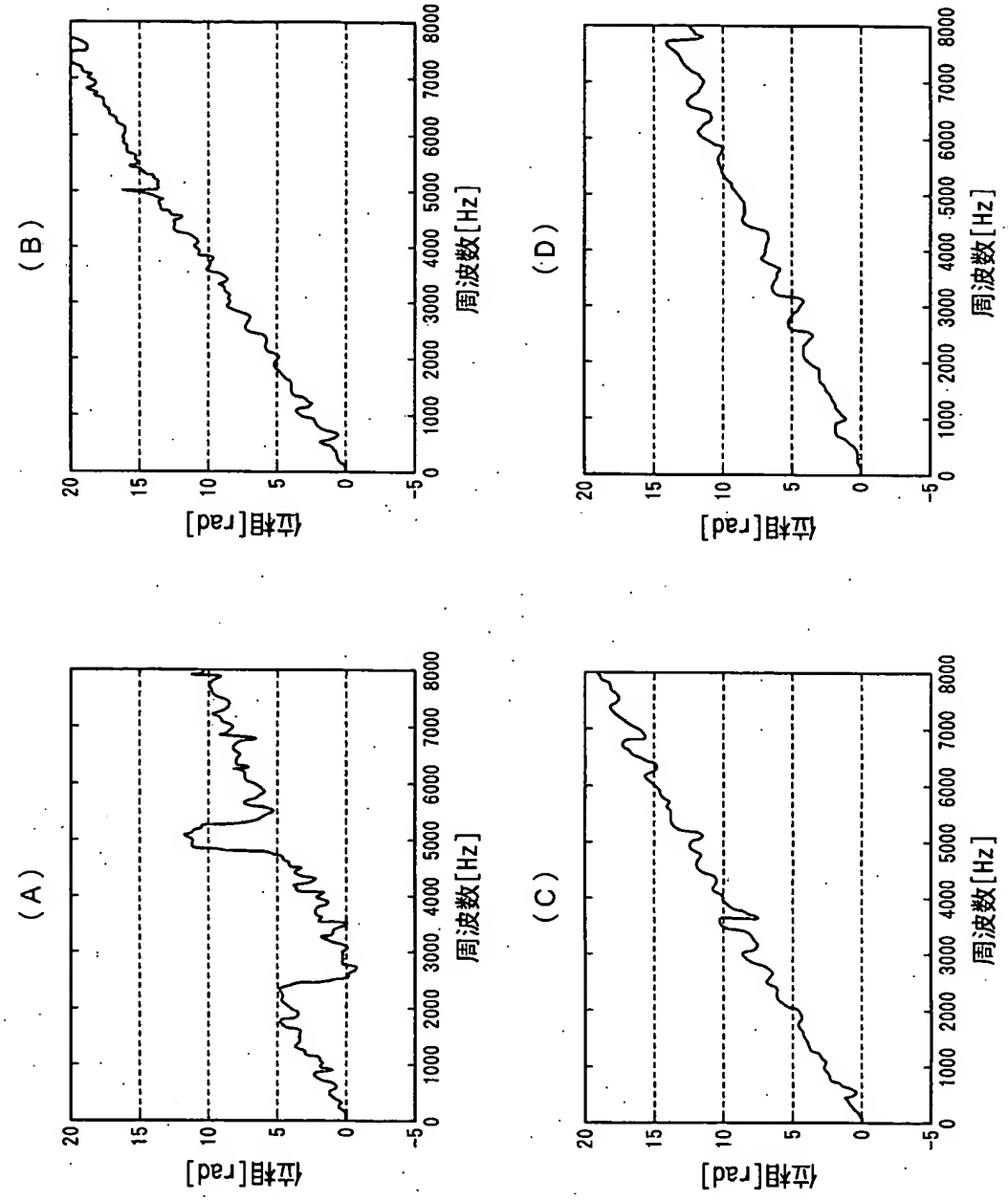
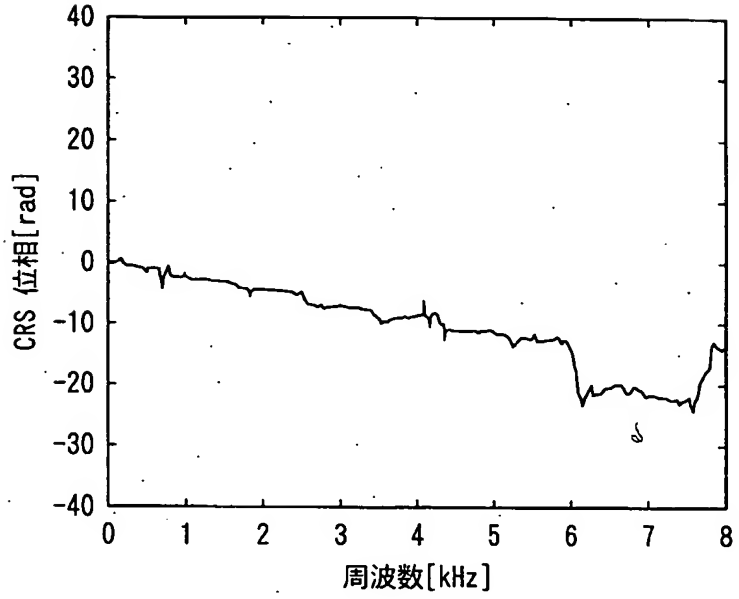

(A)
(B)
(C)
(D)
位相[rad]
周波数[Hz]
20
15
10
5
0
-5
0
1000
2000
3000
4000
5000
6000
7000
8000

図４

(A)


40
30
20
10
0
-10
-20
-30
-40
CRS 位相[rad]
0
1
2
3
4
5
6
7
8
周波数[kHz]

(B)

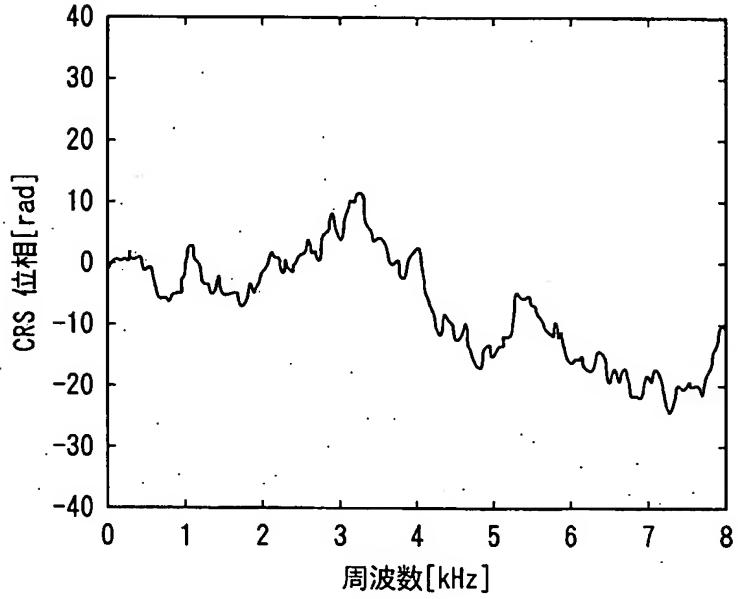
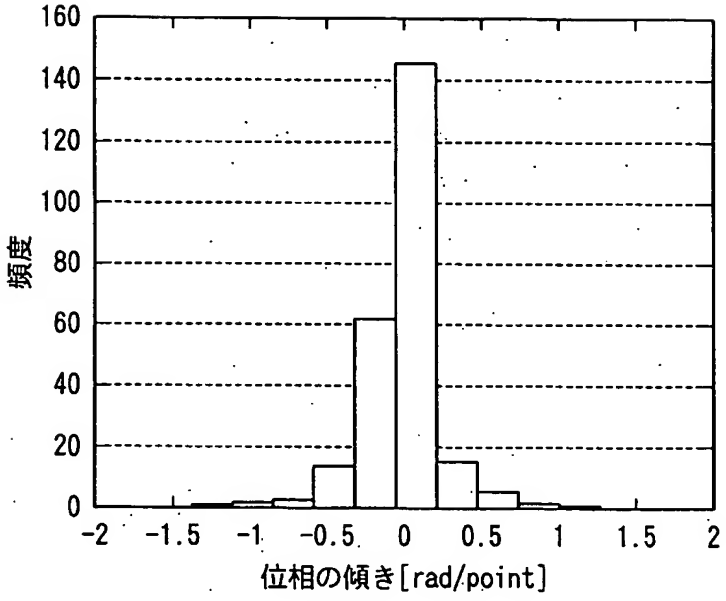

40
30
20
10
0
-10
-20
-30
-40
CRS 位相[rad]
0
1
2
3
4
5
6
7
8
周波数[kHz]

図5

(A)


160
140
120
100
80
60
40
20
0
頻度
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
位相の傾き[rad/point]

(B)

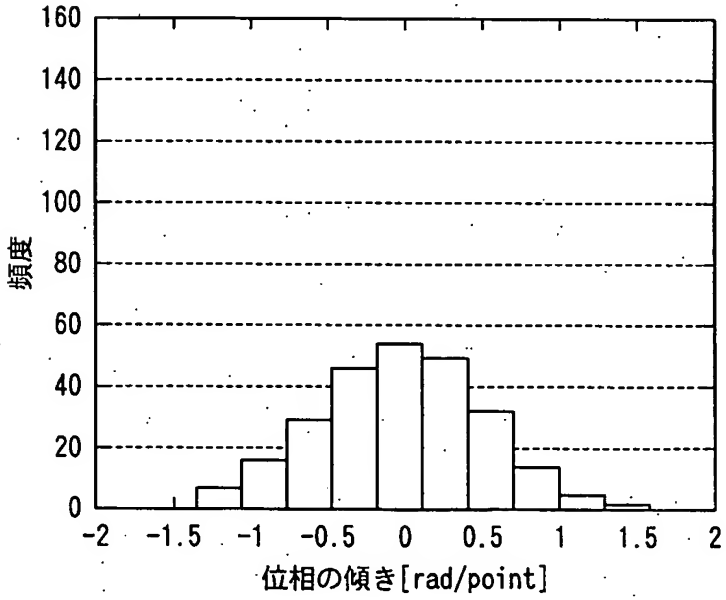

160
140
120
100
80
60
40
20
0
頻度
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
位相の傾き[rad/point]

図6

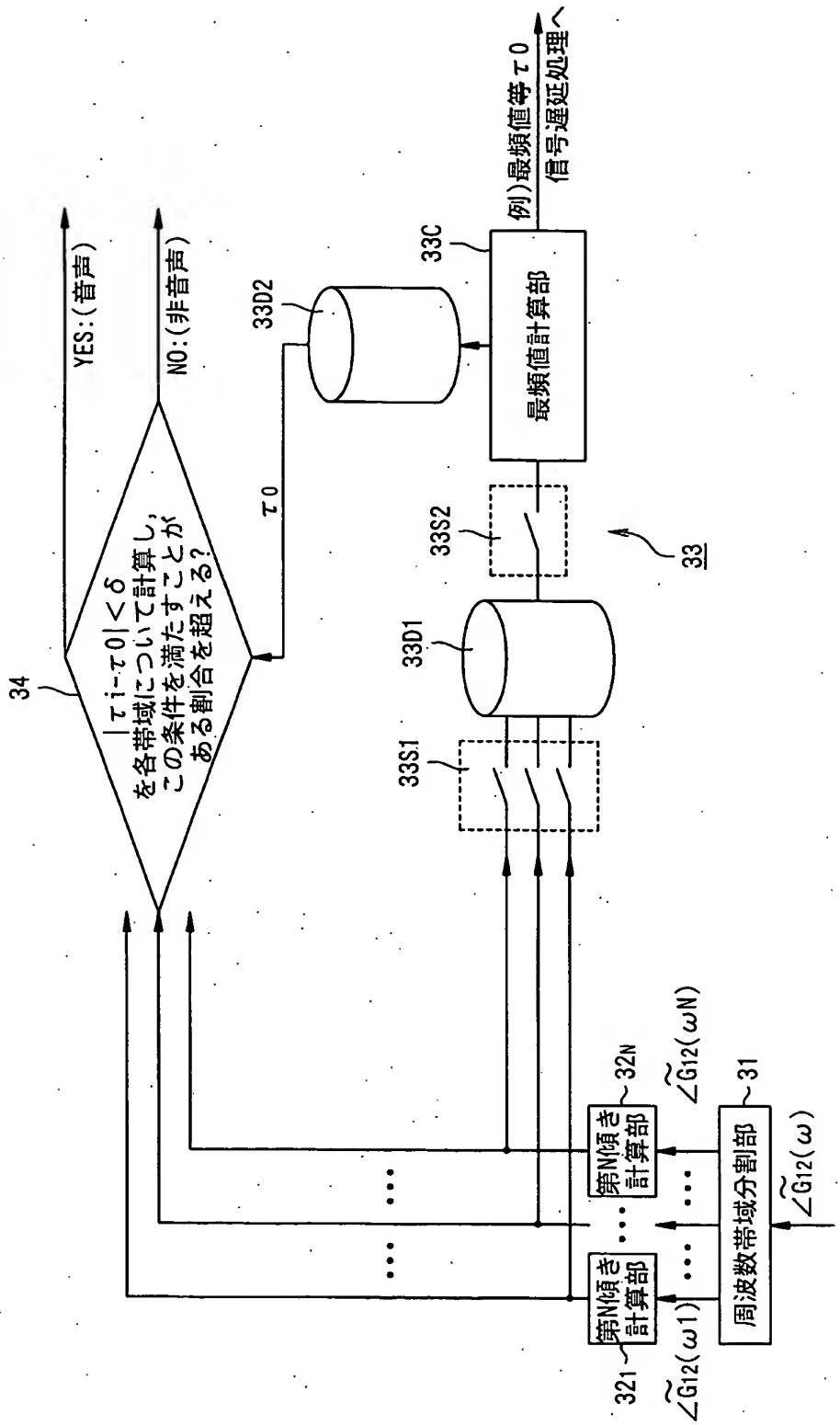

例)最頻値等τ0
信号遅延処理へ
33C
最頻値計算部
33D2
τ0
YES:(音声)
NO:(非音声)
|τi-τ0|<δ
を各帯域について計算し,
この条件を満たすことが
ある割合を超える?
34
33S2
33
33D1
33S1
第N傾き
計算部
32N
第N傾き
計算部
321
周波数帯域分割部
31
∠G̃12(ωN)
∠G̃12(ω1)
∠G̃12(ω)

図７

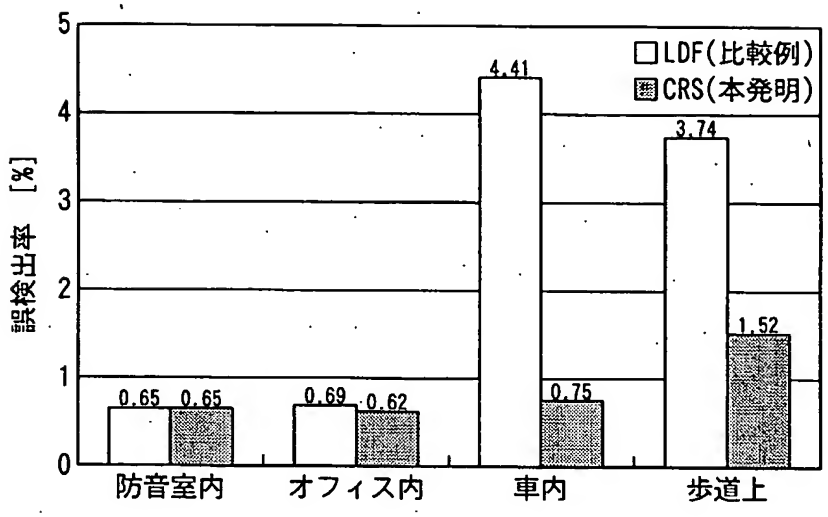

□LDF(比較例)
▨CRS(本発明)
誤出検話率 [%]
5
4
3
2
1
0
0.65
0.65
0.69
0.62
4.41
0.75
3.74
1.52
防音室内
オフィス内
車内
歩道上

図8

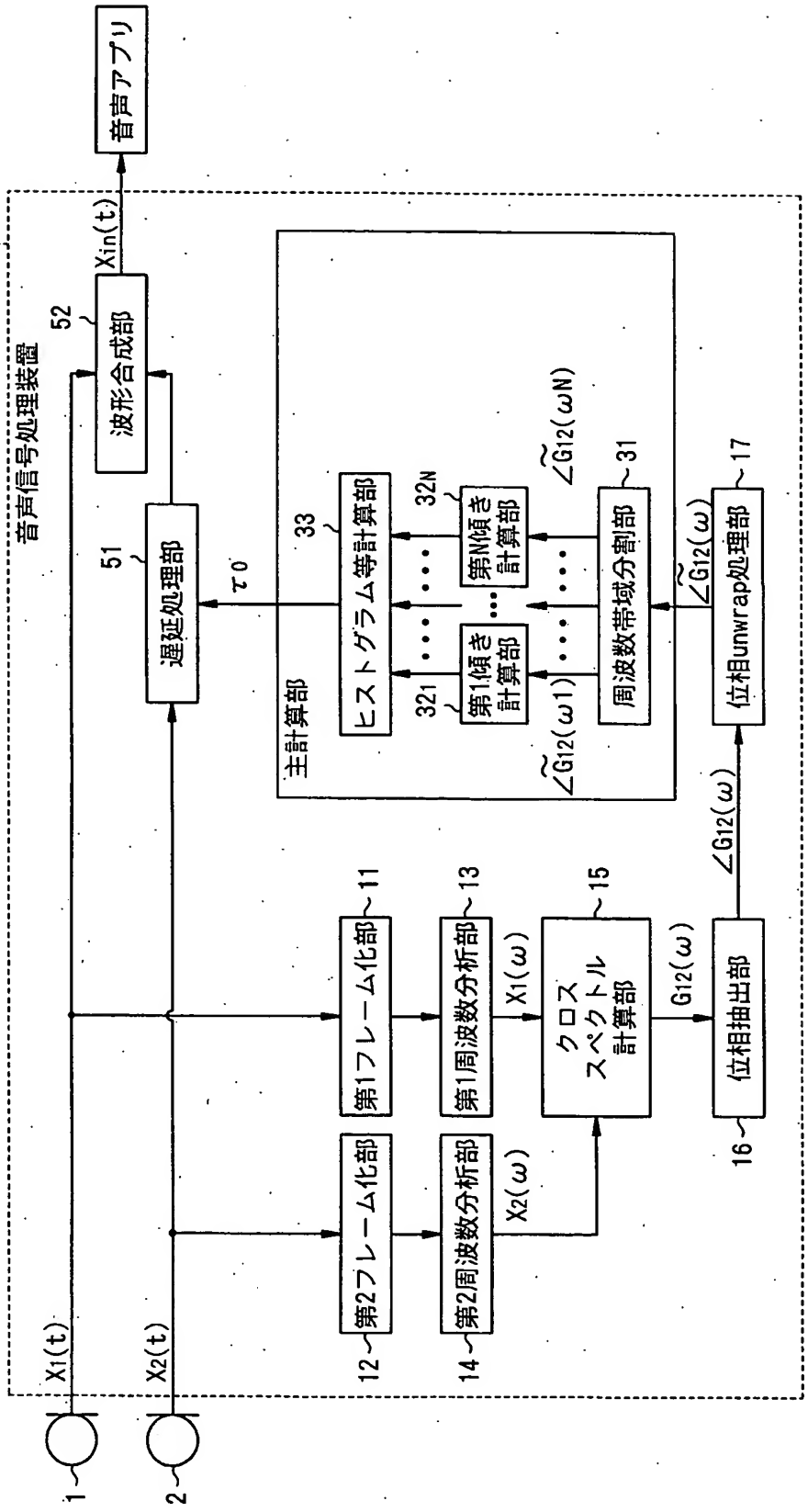

音声信号処理装置
10
音声アプリ
Xin(t)
波形合成部
52
遅延処理部
51
τ0
主計算部
33
ヒストグラム等計算部
321
32N
第1傾き計算部
第N傾き計算部
∠G̃12(ω1)
∠G̃12(ωN)
周波数帯域分割部
31
∠G̃12(ω)
位相unwrap処理部
17
∠G12(ω)
位相抽出部
16
G12(ω)
クロススペクトル計算部
15
X1(ω)
X2(ω)
第1周波数分析部
13
第2周波数分析部
14
第1フレーム化部
11
第2フレーム化部
12
X1(t)
X2(t)
1
2

図９

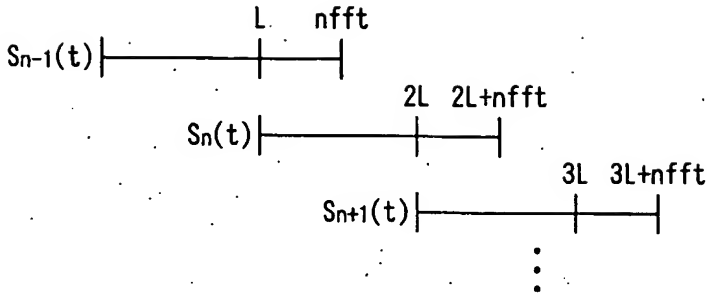

L nfft
Sn-1(t)
2L 2L+nfft
Sn(t)
3L 3L+nfft
Sn+1(t)

図１０

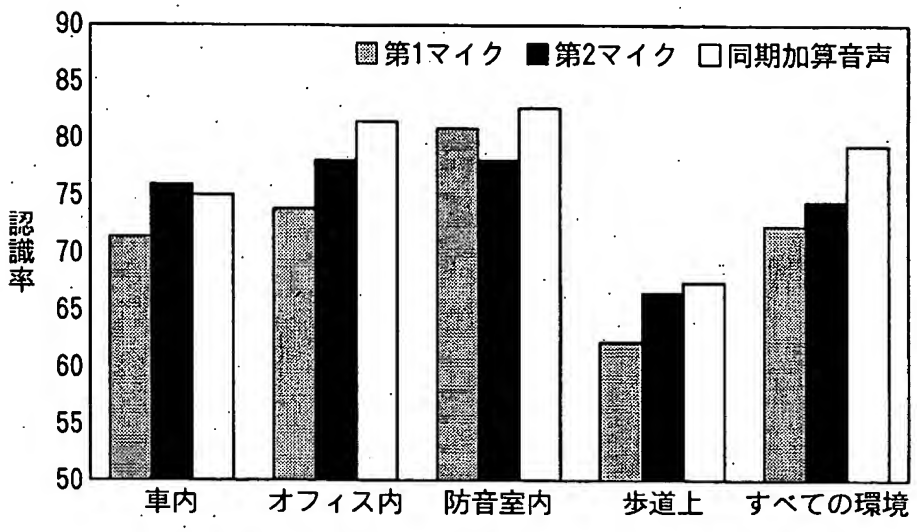

90
85
80
75
70
65
60
55
50
認識率
第1マイク
第2マイク
同期加算音声
車内
オフィス内
防音室内
歩道上
すべての環境

図１１

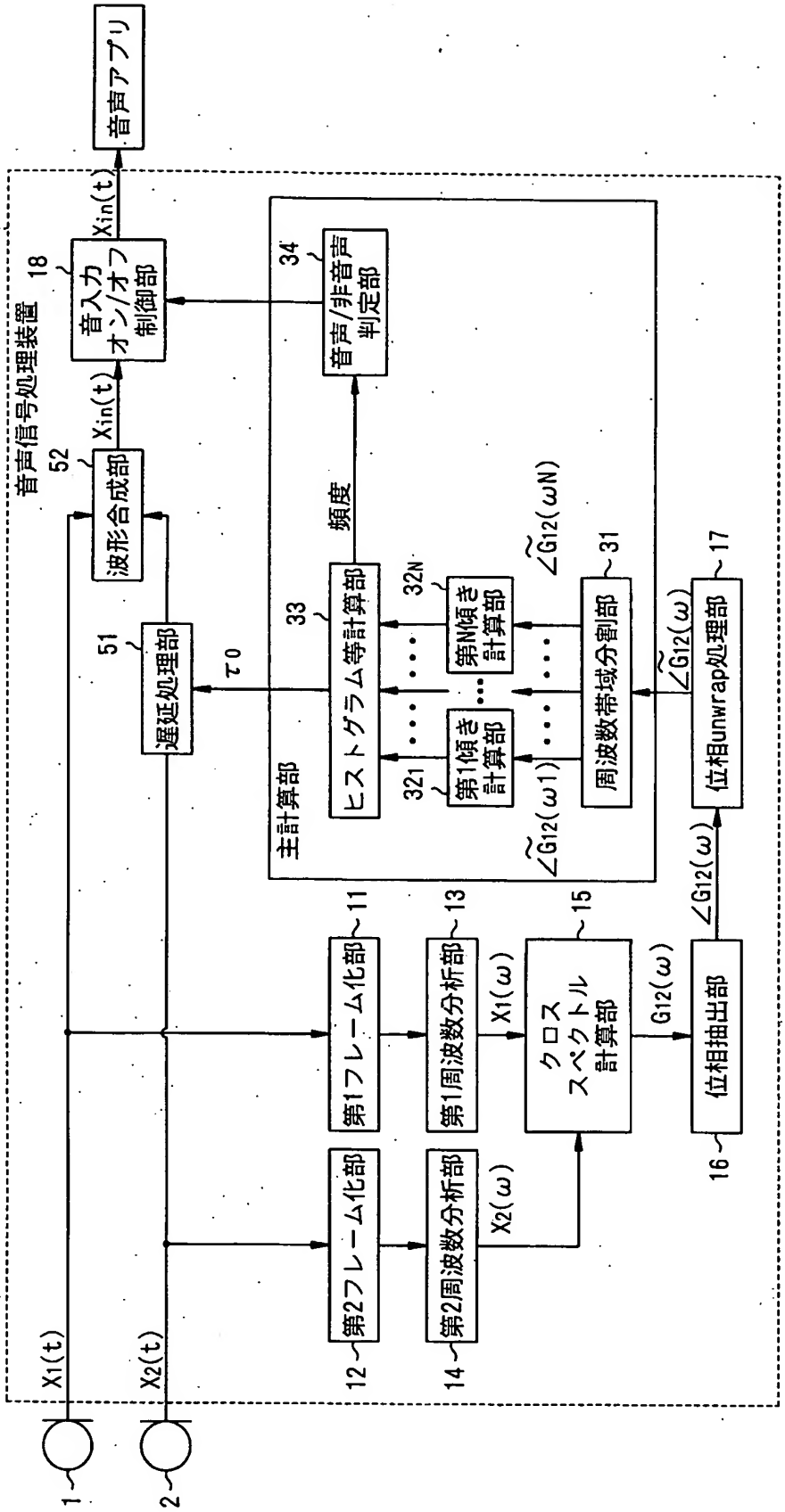
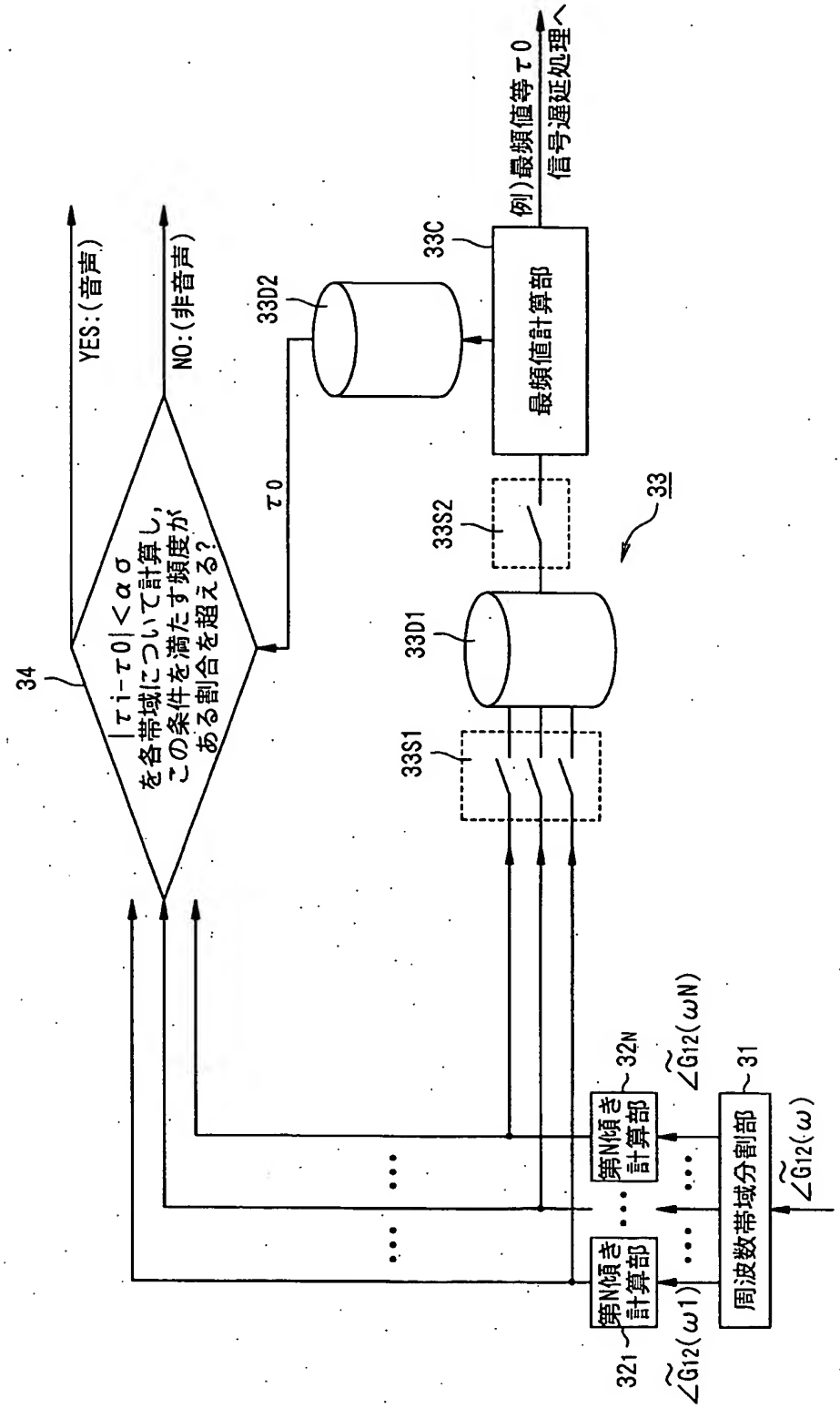

音声アプリ
Xin(t)
10
音声信号処理装置
18
音入力
オン/オフ
制御部
34
音声/非音声
判定部
52
波形合成部
Xin(t)
頻度
51
遅延処理部
τ0
33
主計算部
ヒストグラム等計算部
32N
第N傾き
計算部
321
第1傾き
計算部
∠G̃12(ωN)
∠G̃12(ω1)
31
周波数帯域分割部
∠G̃12(ω)
17
位相unwrap処理部
∠G12(ω)
11
第1フレーム化部
13
第1周波数分析部
X1(ω)
15
クロス
スペクトル
計算部
G12(ω)
位相抽出部
16
X1(t)
X2(t)
12
第2フレーム化部
14
第2周波数分析部
X2(ω)
1
2

図１２


周波数帯域分割部
31
第N傾き計算部
32N
第N傾き計算部
321
33
33S1
33S2
33D1
33D2
33C
最頻値計算部
例）最頻値等τ0
信号遅延処理へ
34
|τi-τ0|＜ασ
を各帯域について計算し，
この条件を満たす頻度が
ある割合を超える？
YES:（音声）
NO:（非音声）
τ0
∠G̃12(ω1)
∠G̃12(ωN)
∠G̃12(ω)